**Preparation/Analysis Big Data - BIA 6305**

**Leonardo Ji**

**Assignment 5**

**Part 1 Hadoop Articles**

* [Top 10 commercial Hadoop platforms](https://www.hadoop360.datasciencecentral.com/blog/top-10-commercial-hadoop-platforms)
* [Hadoop whose to choose](https://www.hadoop360.datasciencecentral.com/blog/hadoop-whose-to-choose)

The first article listed top ten commercial Hadoop vendors: Cloudera, Amazon Web Services, Hortonworks, MapR, IBM, Microsoft HDInsight, Intel distribution for Apache Hadoop, Datastax Enterprise Analytics, Teradata Enterprise Access for Hadoop, Pivotal HD. Only Cloudera, AWS, Hortonworks, Microsoft HDInsight, Pivotal HD offer their own Hadoop distributions while other companies offer the same open source Hadoop distribution or adds-on products. Not surprisingly, Google is not in the Hadoop business. Hadoop ecosystem learned from Google proprietary software like Google MapReduce (Hadoop), Google Bigtable (Hive), Google File System (HDFS).

The second article compare three top Hadoop companies: Cloudera, Hortonworks, and MapR on general information, admin tools, SQL on Hadoop, performance benchmarks.

The second article was written in 2015. Cloudera went public 2017. Its valuation dropped from 4.7 billion as a private company to 2.4 billion as a public company. Hortonworks is still the same valuation of 1.2 billion as a private company. Both Cloudera and Hortonworks’s revenue more than tripled from 2015 to 2018. MapR is still a private company and its valuation is at 950 million.

Hortonworks seems only offers open source software. Hortonworks offers Apache Ambari for admin tools. Cloudera and MapR offer their proprietary Hadoop admin tools: Cloudera Manager, and MapR Control System, respectively.

Hortonworks offers only open source version of Apache Hive as its SQL on Hadoop. Hive is a batch oriented tool using MapReduce in the background. Cloudera offers Impala: a real-time distributed query database. Impala streams results back as it is available, Impala is faster and more interactive than Hive. MapR offers Drill SQL. Drill does not need schema to import data, can take advantage of self-discovering data like XML or JSON, AVRO files. The fourth option: Spark SQL is becoming a more popular Big Data SQL tool. Spark SQL holds data in memory and it is very fast on very large dataset. Spark is an open source software developed by Berkeley. It has many other functionalities like SparkMLlib (machine learning libraries).

Three benchmarks from separate organizations shows Impala is faster than Hive. Spark SQL has advantage over Impala on very large dataset. Hive is more mature software, it supports full set of SQL commands INSERT, UPDATE, DELETE, with full ACID compliance.

In summary, picking Cloudera, Hortonworks, MapR, or Spark is up to organization’s priorities. Hortonworks offers support on open source software, and Hortonworks relies on community driven innovation. Cloudera offers Impala, which is faster more interactive query. MapR offers its own proprietary software on Hadoop platform. Spark SQL is very fast on very large dataset and it is based on open source Spark platform.

**Docker Articles**

* [Docker for Data Science](https://towardsdatascience.com/docker-for-data-science-4901f35d7cf9)
* [How Docker can help you become a more effective data scientist](https://towardsdatascience.com/how-docker-can-help-you-become-a-more-effective-data-scientist-7fc048ef91d5)

Software installation is a pain. The same software worked on my computer might not work in a different versions or libraries. So why use Docker? Docker saves time on installing or updating packages because Docker images contain all packages necessary to run the software. Docker prevents random errors because the same dependencies and libraries are stored in Docker images that runs on any OS platforms. Docker allows anyone to distribute not only software code, but also environment that software requires to run.

Docker is a light way virtual machine that contains everything needed to run a software. Docker does not depend on OS and the same docker containers runs the same way on Windows, Linux, or MacOS. Another benefit is we can send Docker images to Amazon Web Services (AWS) to scale it up when we need more computing resources to run machine learning models.

Dockerfile contains information on how to set the same environments and libraries, retrieve dependencies from DockerHub. DockerHub is an online site that stores docker images. We can push the Docker file to Github repo. The next person would be able to checkout the Dockerfile to reliably run the same code without an issue.

We can create a new docker image from the Dockerfile using the “docker build” command. Docker images is a snapshot of your machine that another person can reproduce the run with the same dependencies and libraries. The "docker image” command lists all Docker images on your computer.

Docker container is a running instance of Docker image. If your Docker image is a web server, Docker container is an instance of web server running on your computer. We can use the “docker load” to load the image then the “docker run” command to create container from the image. The “docker ps” command lists all running Docker containers on your computer.

We can interact with Docker container with “docker exec” commands. We can save the snapshot of container to a new Docker image with “docker commit” command. Then we can use “docker push” command push your image to remote Dockerhub. A new Docker images is born. Someone else can download it and use it.

In summary to be an effective Data Scientist, I can use Docker to share reproducible results; publish work as a Docker container so that anyone can reproduce it; move my container to AWS or Azure when larger computing resources are needed; make change to customized settings or libraries in a Dockerfile, and it would always run the same way next time; use someone else’s Docker image to run complicated code like TensorFlow.